Understanding Data Augmentation in Neural Machine Translation: Two Perspectives towards Generalization

Guanlin Li$^{\epsilon}$ [*], Lemao Liu$^{\lambda}$, Guoping Huang$^{\lambda}$, Conghui Zhu$^{\epsilon}$, Tiejun Zhao$^{\epsilon}$, Shuming Shi$^{\lambda}$
$^{\epsilon}$Harbin Institute of Technology, $^{\lambda}$ Tencent AI Lab
{epsilonlee.green}@gmail.com, {chzhu, tjzhao}@hit.edu.cn,
{redmondliu, donkeyhuang, shumingshi}@tencent.com

Abstract

Many Data Augmentation (DA) methods have been proposed for neural machine translation. Existing works measure the superiority of DA methods in terms of their performance on a specific test set, but we find that some DA methods do not exhibit consistent improvements across translation tasks. Based on the observation, this paper makes an initial attempt to answer a fundamental question: what benefits, which are consistent across different methods and tasks, does DA in general obtain? Inspired by recent theoretic advances in deep learning, the paper understands DA from two perspectives towards the generalization ability of a model: input sensitivity and prediction margin, which are defined independent of specific test set thereby may lead to findings with relatively low variance. Extensive experiments show that relatively consistent benefits across five DA methods and four translation tasks are achieved regarding both perspectives.

1 Introduction

Data Augmentation (DA) is a training paradigm that has been proved to be very effective in many modalities (Park et al., 2019; Perez and Wang, 2017; Sennrich et al., 2016a), especially for classification (Perez and Wang, 2017). In structured domain, Neural Machine Translation (NMT) is the frontier of DA research (Sennrich et al., 2016a; Norouzi et al., 2016; Zhang and Zong, 2016; Fadaee et al., 2017; Wang et al., 2018; Zhang et al., 2019; Edunov et al., 2018; Fadaee and Monz, 2018). However, by investigating a variety of DA methods, we find that their test performance across different translation tasks does not exhibit consistent improvement, and this phenomenon can be initially observed in (Wang et al., 2018) as well. The reason might be the evaluation metric on a specific test set when compared to the whole data population, which generates all possible data, has large variance so that leads to the inconsistency. This evaluation dilemma is also recognized and explored by Recht et al. (2018, 2019); Werpachowski et al. (2019), and is especially notorious for language generation tasks (Chaganty et al., 2018; Hashimoto et al., 2019) where the evaluation metrics, e.g. BLEU (Papineni et al., 2001), are extrinsic and heavily relies on the reference provided. Therefore, we ask a fundamental question: what benefits, which are more consistent across different DA methods and translation tasks, can DA in general obtain?

A direct answer to the above question is to use generalization gap (Kawaguchi et al., 2018) defined by the difference between population risk and empirical risk. This measure does not rely on any specific test set, accurately depicts generalization but is intractable to compute. So recently, many theorists have proposed either non-vacuous generalization bound (Dziugaite and Roy, 2017; Zhou et al., 2019) or novel generalization measures (Novak et al., 2018; Bartlett et al., 2017; Neyshabur et al., 2017; Jiang et al., 2019) to roughly reflect the gap. Inspired by them, we propose to understand the benefits of DA from two perspectives: input sensitivity and prediction margin. The proposed underlying two measures are well adapted from Novak et al. (2018) and Bartlett et al. (2017) and can be computed only on the train samples to unveil the consistent benefits of DA. Under a carefully designed fair setting over four different translation tasks, we examine five methods from two main categories of DA and compare them with a model trained without DA. The empirical experiments demonstrate the following findings: a). DA methods exhibit more consistent effects across different translation tasks in terms of both measures. b). DA methods can either allevi-

[*] Work done at Tencent AI Lab.

ate input sensitivity or promote prediction margin. By and large, our main contributions are:

- We make an initial attempt to understand the essence of DA in NMT by investigating its benefits which are relatively consistent across five DA methods and four translation tasks.
- We highlight two perspectives towards generalization to measure the benefits of DA in NMT and study them with carefully designed fair experiments.

2 DA Methods in NMT

2.1 Training Objective Decomposition

Given the train set $\mathcal{T}$ the baseline NMT model $p_\theta(\mathrm{y}|\mathrm{x})$ without using DA is trained under the empirical data distribution $\hat{p}(X, Y|\mathcal{T})$ through maximum likelihood estimation:

$$J_{\text{MLE}} = \mathbb{E}_{\mathrm{x,y}\sim\hat{p}(X,Y|\mathcal{T})}[\log p_\theta(\mathrm{y}|\mathrm{x})], \qquad (1)$$

where $\hat{p}$ is a mixture of Dirac distribution concentrated around each training instance with uniform mixture coefficients ($1/|\mathcal{T}|$). Then we define the augmentation (AUG) model as a conditional distribution over the train set, $q(X, Y|\mathcal{T})$. [1] Under the AUG model, the training objective becomes:

$$J_{\text{AUG}} = \mathbb{E}_{\mathrm{x,y}\sim q(X,Y|\mathcal{T})}[\log p_\theta(\mathrm{y}|\mathrm{x})]. \qquad (2)$$

More realistically, for any DA method in any training run, we can collect the augmented instances to form a set $\mathcal{A}$ distinguishing $\mathcal{T}$, when considering the curriculum of mixing $\mathcal{A}$ with the original train $\mathcal{T}$. Since we would like to derive a conceptual framework that reflects this form of importance weighting, we further decompose AUG model into a linear interpolation (α) of $\hat{p}(X, Y|\mathcal{T})$ and an augmentation distribution $q_{\text{AUG}}(X, Y|\mathcal{T})$:

$$\alpha \cdot \hat{p}(X, Y|\mathcal{T}) + (1 - \alpha) \cdot q_{\text{AUG}}(X, Y|\mathcal{T}), \qquad (3)$$

where α controls the mixture ratio within a batch during SGD training. The ratio has been founded as an important factor influencing final performance (Sennrich et al., 2016a; Fadaee et al., 2017; Edunov et al., 2018; Fadaee and Monz, 2018).

Method	Fr⇒En	En⇒Fr	Zh⇒En	En⇒De
Baseline	38.38 (5)	38.88 (6)	17.25 (6)	26.19 (4)
RAML	*+0.22* (3)	*+0.67* (3)	*+0.23* (4)	*-0.16* (6)
SO	*+0.01* (4)	*+0.62* (4)	*+0.02* (5)	*-0.15* (5)
ST	*-0.13* (6)	*+0.46* (5)	*+1.51* (2)	*+0.83* (2)
TA	*+0.62* (2)	*+1.13* (1)	*+2.41* (1)	*+1.01* (1)
BT	*+0.82* (1)	*+0.99* (2)	*+1.06* (3)	*+0.39* (3)

Table 1: Main BLEU results (CTC=0.62).

Key factors Through Eq. 3, we can identify two key factors for conducting fair experiments: a) the number of SGD updates on every original training instance means how much the model learns from $\mathcal{T}$; b) the mixture ratio means how much the model learns from $\mathcal{A}$ online, with which together balance the learning of the translation knowledge.

2.2 Settings and Main Performance

Settings By carefully controlling the above two factors, we conduct fair and extensive experiments with Transformer (Vaswani et al., 2017) on four translation tasks for five DA methods. Fairseq (Ott et al., 2019) is used as our codebase. We use standard benchmarks IWSLT17 En-Fr, WMT19 Zh-En, WMT19 En-De, where we train both translation directions on the IWSLT corpus. The five DA methods are briefly summarized as follows:

- RAML: reward-augmented maximum likelihood training, which augment the target-side with a sampling distribution $P(Y|Y^*)$ concentrated around Y^* (Norouzi et al., 2016).
- Switchout (SO): similar to RAML, but also adds the some kind of augmentation to the source-side (Wang et al., 2018).
- Self-training (ST): fix the source-side, uses an forward NMT model to generate the target-side (Zhang and Zong, 2016).
- Target-agree (TA): similar to ST, but uses a forward NMT model with right-to-left decoder (Zhang et al., 2019).
- Back-translation (BT): fix the target-side, uses an backward NMT model to generate the source-side (Sennrich et al., 2016a).

The implementation of RAML and SO are borrowed from the Appx. of Wang et al. (2018). [2]

[1] In the paper, we do not consider using monolingual data for DA thus conditioning only on bilingual data since this will further bring monolingual data selection discussed in Fadaee and Monz (2018) as a factor to influence the performance of different DA methods; we leave this factor for future study.

[2] We categorize and analyze how we choose or train DA methods with a generative formulation in Appx. A and B.

Method	Fr⇒En	En⇒Fr	Zh⇒En	En⇒De
Baseline	0.565 (6)	0.623 (4)	0.422 (6)	0.682 (5)
RAML	*-0.056* (2)	*+0.003* (5)	*-0.008* (4)	*-0.003* (4)
SO	*-0.082* (1)	*-0.099* (1)	*-0.053* (2)	*-0.143* (1)
ST	*-0.034* (3)	*-0.009* (3)	*-0.054* (1)	*-0.116* (2)
TA	*-0.023* (5)	*-0.010* (2)	*-0.043* (3)	*-0.013* (3)
BT	*-0.026* (4)	*+0.035* (6)	*-0.007* (5)	*+0.232* (6)

Table 2: Sensitivity measure (CTC=0.72).

CTC Table 1 shows the main BLEU results of different methods on the test set. However, we cannot identify the best DA method because their rankings across the four translation tasks vary a bit. To measure the degree of consistency, we use a correlation measure called Kendall's coefficient of concordance (Kendall and Smith, 1939; Mazurek, 2011) to evaluate the correlation of the rankings produced on the four translation tasks (appx. C). The value shows strong consistency (correlation) of different rankings when it is close to 1. We call the correlation value Cross-Task Consistency measure or CTC. The CTC for the BLEU measure is 0.62, which is of weak consistency. This phenomenon might be a result of the intrinsic nature of using a single specific test as a substitute of the whole data population for evaluation. In the next section, we introduce two measures that are more consistent (with close-to-1 CTC value). They in some extent reflect the model generalization and are easy-to-compute as well.

3 Two Measures Towards Generalization

We attempt to understand the benefits that DA can obtain through the quantification of input sensitivity and prediction margin. The two measures are adapted from Novak et al. (2018) and Bartlett et al. (2017); Neyshabur et al. (2017); Jiang et al. (2019). They have been proved through massive experiments to be correlated with model generalization. Our main purpose here is to utilize them to unveil the consistency property (measured by CTC) of DA across different methods and translation tasks. The next two subsections define the two measures and report their statistics on subsamples of the train set respectively.

3.1 Input Sensitivity

Input sensitivity is the sensitivity of the loss computed from the model towards a minor change of input representation. Given a point of interest x, the original form in Novak et al. (2018) is defined as the expected Jacobian norm of the loss vector $\log p_\theta(\cdot|\mathrm{x})$ and p_θ is a softmax classifier:

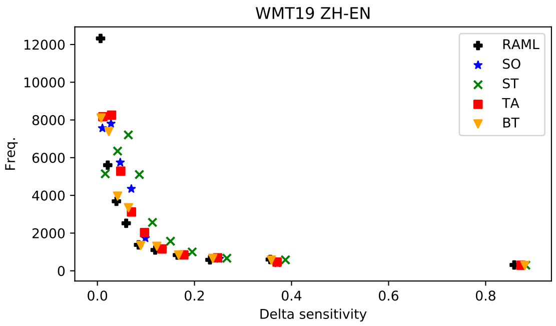

Figure 1: Δ sensitivity binned avg. token freq. statistics. Each point represents a bin from which we compute the token level average Δ sensitivity between that DA method and the baseline and the token level average frequency as its x and y coordinate value.

$$\mathbb{E}_{\mathrm{x}}||J(\mathrm{x})||_F, \tag{4}$$

where $J(\mathrm{x}) = \partial \log p_\theta(\cdot|\mathrm{x})/\partial \mathrm{x}^T$, and $||\cdot||_F$ the Frobenius or L2 norm of the matrix. The paper also suggests a more predictive quantity of the generalization ability. That is to take advantage of the label y of x and only compute the L2 norm of a slice of the Jacobian matrix indexed by the label. We adopt the later measure which is the gradient norm of the loss scalar indexed by y to x:

$$\mathbb{E}_{\mathrm{x}}||J(\mathrm{x})_y||_F. \tag{5}$$

If $\mathrm{x} \in \mathbb{R}^d$ lies in a space with differential structure, we can apply Eq. 5 directly. But in NMT the naive representation of an instance (x, y) is the token index given by the vocabulary, so we cannot compute the gradient of the loss with respect to x. We follow Sundararajan et al. (2017) and use the result of x after embedding lookup as its learned representation, denoted as $\mathrm{Emb}(\mathrm{x}) \in \mathbb{R}^{L_{\mathrm{x}} \times d}$ where L_{x} is the length of the input and d the size of the embedding. By regarding the translation model $p_\theta(\mathrm{y}|\mathrm{x})$ as a function that decomposes at each step of y given $\mathrm{Emb}(\mathrm{x})$ as input to get a scalar average log likelihood, denoted as $\mathcal{L}_{\mathrm{x,y}} = \frac{1}{L_{\mathrm{y}}} \sum_t \log p_\theta(\mathrm{y}_t|\mathrm{y}_{<t}, \mathrm{x})$. Moreover, initial experiments on just using the single original x_i to evaluate the gradient will still result in inconsistency, due to the non-equivalence of the localness concept compared with the continuous setting, i.e. for language input, the localness is between discrete inputs in the neighbor of x_i. So we

evaluate the sensitivity of x_i by averaging gradient norms over its k nearest neighbor $\mathrm{x}_{i(j)} \in \mathrm{kNN}[\mathrm{x}_i]$ through cosine similarity between word embeddings. We set k to 5 in our experiment to guarantee words in the k nearest neighbor has similar semantic meaning. Formally, we define the input sensitivity of an NMT model as:

$$\mathbb{E}_{(\mathrm{x},\mathrm{y})} \frac{1}{L_{\mathrm{x}}} \sum_{i} \frac{1}{k} \sum_{\mathrm{x}_{i(j) \in \mathrm{kNN}[\mathrm{x}_i]}} \left|\left| \frac{\partial \mathcal{L}_{\mathrm{x},\mathrm{y}}}{\partial \mathrm{Emb}(\mathrm{x})}_{i(j)} \right|\right|_F, \tag{6}$$

where the $\mathrm{Emb}(\mathrm{x})_i$ is the embedding lookup of the i^{th} token index, so we compute the average token-wise gradient of each instance.

We use subsamples of the train set to approximately compute the expectation in Eq. 6 and the overall statistics are shown in Table 2. Similar to Table 1, the DA methods are shown in their Δ value respect to the baseline. A first thing to notice is that the ranking is more steady across tasks (CTC=0.72). It also shows that for input x, DA in general can reduce the gradient norm of the prediction loss on $\mathrm{Emb}(\mathrm{x})_i$, which shows that DA can obtain more stable model towards data corruption.

To further understand what effect DA in general has on each input token type, we compute the Δ sensitivity between the baseline and one DA method on the same token type and sort them according to the Δ with positive value (which means DA reduces the sensitivity of that token type). We then divide the sorted types into ten bins and compute the average token type frequency of that bin. As shown in Figure 1, DA in general, improves the sensitivity of token types with relatively low frequency more than those with high frequency, thus may somehow improve the translation quality of low frequency token types.

3.2 Prediction Margin

Margin is a classic concept in support vector machine (Vapnik, 2013), which is defined as the geometric distance between the support vectors and the decision boundary. Larger margin implies better generalization. In nonlinear case, it reflects the distance of a correctly classified input representation with class i to move towards the decision boundary between i and any other class j (Jiang et al., 2019). However, since the decision boundary does not have analytical form due to nonlinearity, computing the geometric distance is intractable. In our setting we regard NMT model as doing step-wise classification with $\mathrm{z} = (\mathrm{x}, \mathrm{y}_{<t})$ as input feature and y_t as the label. In Bartlett et al. (2017), the original definition of the margin of correctly predicted input is:

Method	**Fr⇒En**	**En⇒Fr**	**Zh⇒En**	**En⇒De**
Baseline	0.797 (4)	0.592 (4)	0.756 (4)	0.679 (4)
RAML	*-0.028* (6)	*-0.063* (6)	*-0.024* (6)	*-0.006* (5)
SO	*-0.027* (5)	*-0.060* (5)	*-0.022* (5)	*-0.009* (6)
ST	*+0.046* (1)	*+0.036* (1)	*+0.035* (1)	*+0.044* (1)
TA	*+0.037* (2)	*+0.019* (2)	*+0.025* (2)	*+0.043* (2)
BT	*+0.017* (3)	*+0.015* (3)	*+0.004* (3)	*+0.016* (3)

Table 3: Margin measure (CTC=0.98).

$$\frac{p_\theta(\mathrm{y}_t|\mathrm{z}) - max_{v' \neq \mathrm{y}_t} p_\theta(v'|\mathrm{z})}{\mathcal{R} \cdot ||\mathrm{x}||_2/N}, \tag{7}$$

where y_t is the ground-truth label, v' another class type, $\mathcal{R}$ the spectral complexity of the model and N the number of training instance in the train subsamples for computing their margins. We simplify Eq. 7 to only consider the numerator. The reasons are: a) under the same model architecture, $\mathcal{R}$s are very close across different DA methods; b) we can omit $||x||_2/N$ since it remains unchanged as well. In this way, we can map every z, y_t to a margin with label type $\mathrm{y}_t = v$, where $v \in \mathcal{V}_{tgt}$:

$$m^v_{\mathrm{z},\mathrm{y}_t} := p_\theta(\mathrm{y}_t|\mathrm{z}) - max_{v' \neq \mathrm{y}_t} p_\theta(v'|\mathrm{z}). \tag{8}$$

So for every target token type v we can collect a set of margins $\{m^v_{\mathrm{z},\mathrm{y}_t}\}$, and the margin sets of all token types are combined as the total margin set $\cup_v \{m^v_{\mathrm{z},\mathrm{y}_t}\}$. Following Neyshabur et al. (2017), we do not compute the minimum margin of the total set which can be highly sensitive to outliers. Instead, if the total set has cardinality N', we obtain the $\epsilon N'$-th smallest margin from the set as the overall prediction margin, with a tolerant coefficient $\epsilon \in [0, 0.1]$ (ϵ is set to 0.001). [3] We can also obtain token-wise prediction margin from $\{m^v_{\mathrm{z},\mathrm{y}_t}\}$, which is the prediction margin of a specific token type v.

The overall prediction margins are listed in Table 3. The relative rank is highly consistent across the four translation tasks (CTC=0.98). Although RAML and SO seem to be inferior to the baseline, other DA methods improve the margin in general. We give a possible explanation for this in the next subsection. Similar to the previous subsection, we

[3] Note that, we have tried different tolerant coefficents and find they do not effect the final rankings of DA methods.

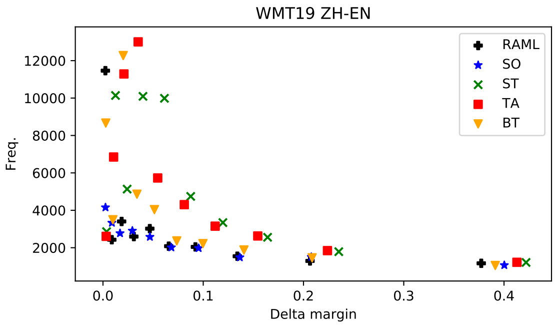

Figure 2: Δ margin binned avg. token freq. statistics.

also report the average token type frequency of each margin binned token groups in Figure 2 and find that DA, in general, brings larger margin improvement over low frequency tokens.

3.3 Discussion

Why use these two measures? We have conducted a relatively complete survey of the recent measures towards measuring generalization ability proposed by the deep learning community, such as model complexity (Zhang et al., 2016; Neyshabur et al., 2017), flatness (Dinh et al., 2017), stiffness (Fort et al., 2019) and second order Hessian of the input or the number of linear regions in hiddens (Novak et al., 2018; Montufar et al., 2014). Some of those measures have complex definition such as linear regions, others are very expensive to compute for models as large as Transformer such as Hessian and stiffness. However, we compute weight norm with different forms proposed in Neyshabur et al. (2017) and find no regularity which suggests that the complexity measure through norms for networks architectures like Transformer or convolutional/recurrent neural networks might be very different from simple feed-forward ones which might be still an open problem in theoretic deep learning community. As a matter of fact, due to computational easiness and large-scale empirical evidence, we choose sensitivity and margin the measures.

Why no absolute consistency between the two measures? In Section 3.1 and 3.2, the two measures do not show well consistency between them: under the margin based measure, RAML and SO do not exhibit superiority over the baseline like they do under the sensitivity measure. One reason might be: despite the measures are empirically proved to reflect generalization, they are only one view towards generalization respectively. Specifically, in recent generalization theory (Novak et al., 2018; Bartlett et al., 2017), the measures are evaluated between models with extremely evident difference in generalization ability (measured by test performance difference), for example, between models trained with random labels and true labels. Instead, our comparison is among models with similar capacity and are well-trained, which rises challenge for us to get very consistent statistics through a single view. This may inspire us to combine multiple views of model training to design better measures with stronger correlation.

4 Conclusion and Future Work

This paper aims at delivering relatively consistent benefit measures of DA due to the phenomenon of inconsistant BLEU improvement across translation tasks. To our expect, the proposed two measures exhibit relative consistency (especially prediction margin) on five DA methods across four translation tasks, which demonstrate that DA can benefit model with improved sensitivity or prediction margin especially for low frequency words.

However, the problem of intrinsic evaluation or better understanding of the unreasonable effective of DA should just be a start. DA is a trade-off between noise vs. knowledge injection, so it could be a nice theoretic direction to think about DA under statistical query model (Kearns, 1998) with translation between formal languages (ws-, 2019). This could inspire another essential question: what is the intrinsic properties of the augmented data (Branchaud-Charron et al., 2019) that matter in discrete domain. Applications like active data selection (Coleman et al., 2019) guided with margin or sensitivity can be derived. In general, understanding NMT model's behavior (not only with DAs) beyond BLEU (Neubig et al., 2019) should be taken seriously, e.g. to design a bevavior suite like Osband et al. (2019) is most valuable.

Acknowledgements

The authors are grateful to all the anonymous reviewers for pointing out the topic of intrinsic evaluation of DA methods to summarize our work as well as proposing several critical issues regarding limitations of the proposed measures. We thank Xintong, Zhirui for valuable discussions; Xintong, Chaoqun, Yong Jiang for proof-reading. Tiejun Zhao is supported by National Key R&D Program of China (Project No. 2017YFB1002102).

References

2019. *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*. Association for Computational Linguistics, Florence.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. 2017. Spectrally-normalized margin bounds for neural networks. In *NIPS*.

Frederic Branchaud-Charron, Andrew Achkar, and Pierre-Marc Jodoin. 2019. Spectral metric for dataset complexity assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3215–3224.

A. Chaganty, S. Mussmann, and P. Liang. 2018. The price of debiasing automatic metrics in natural language evaluation. In *Association for Computational Linguistics (ACL)*.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. *CoRR*, abs/1606.04596.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. 2019. Selection via proxy: Efficient data selection for deep learning. *ArXiv*, abs/1906.11829.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. 2017. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1019–1028. JMLR. org.

Gintare Karolina Dziugaite and Daniel M. Roy. 2017. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *CoRR*, abs/1703.11008.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *EMNLP*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *ACL*.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *EMNLP*.

Stanislav Fort, Paweł Krzysztof Nowak, and Srini Narayanan. 2019. Stiffness: A new perspective on generalization in neural networks. *arXiv preprint arXiv:1901.09491*.

Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *ICML*.

T. Hashimoto, H. Zhang, and P. Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *North American Association for Computational Linguistics (NAACL)*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. 2019. Predicting the generalization gap in deep networks with margin distributions. *CoRR*, abs/1810.00113.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. 2018. Generalization in deep learning. *CoRR*, abs/1710.05468.

Michael Kearns. 1998. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006.

Maurice G Kendall and B Babington Smith. 1939. The problem of m rankings. *Annals of mathematical statistics*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *EMNLP*.

Jiří Mazurek. 2011. Evaluation of ranking similarity in ordinal ranking problems. *Acta academica karviniensia*, 2:119–128.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. 2014. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL) Demo Track*, Minneapolis, USA.

Behnam Neyshabur, Srinadh Bhojanapalli, David A. McAllester, and Nathan Srebro. 2017. Exploring generalization in deep learning. In *NIPS*.

Mohammad Norouzi, Samy Bengio, Zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. 2016. Reward augmented maximum likelihood for neural structured prediction. In *NIPS*.

Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. 2018. Sensitivity and generalization in neural networks: an empirical study. *CoRR*, abs/1802.08760.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepezvári, Satinder Singh, Benjamin Van Roy, Richard S. Sutton, David Silver, and Hado van Hasselt. 2019. Behaviour suite for reinforcement learning. *ArXiv*, abs/1908.03568.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Luis Perez and Jason Wang. 2017. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2018. Do cifar-10 classifiers generalize to cifar-10? *CoRR*, abs/1806.00451.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? *CoRR*, abs/1902.10811.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *ICML*.

Vladimir Vapnik. 2013. *The nature of statistical learning theory*. Springer science & business media.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Stefan Wager, Sida I. Wang, and Percy S. Liang. 2013. Dropout training as adaptive regularization. In *NIPS*.

Sida I. Wang, Mengqiu Wang, Stefan Wager, Percy S. Liang, and Christopher D. Manning. 2013. Feature noising for log-linear structured prediction. In *EMNLP*.

Xinyi Wang, Hieu Quang Pham, Zihang Dai, and Graham Neubig. 2018. Switchout: an efficient data augmentation algorithm for neural machine translation. In *EMNLP*.

Roman Werpachowski, András György, and Csaba Szepesvári. 2019. Detecting overfitting via adversarial examples. *CoRR*, abs/1903.02380.

Yingce Xia, Di He, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *NIPS*.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Daniel Jurafsky, and Andrew Y. Ng. 2017. Data noising as smoothing in neural network language models. *CoRR*, abs/1703.02573.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *ArXiv*, abs/1611.03530.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *EMNLP*.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *AAAI*.

Zhirui Zhang, Shuangzhi Wu, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2019. Regularizing neural machine translation by target-bidirectional agreement. *CoRR*, abs/1808.04064.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. 2019. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. In *ICLR 2019*.

A Categorization of DA Methods

By characterizing the second term in Eq. 3, recent DA methods can be conceptually summarized in Table 4. In RAML, given an instance from the empirical data distribution $\hat{p}$, the augmented tgt is sampled independently without considering the equivariance of the src $\tilde{\mathrm{x}} = \mathrm{x}$ as well as its linguistic smoothness. The guarantee of not incurring much noise is the concentration property of $p_{\tilde{\mathrm{x}}|\mathrm{x}}$ around x. SO modifies RAML to consider augmenting the src as well. This kind of noise injection paradigm has been previously studied by Wager et al. (2013); Wang et al. (2013); Xie et al. (2017) as feature/data noising, and can be seen as a kind of regularization technique as Dropout (Srivastava et al., 2014), also pointed out in Wang et al. (2018). So we call them data noising (dn) based DA methods.

For the ST, TA, BT and DA4Low methods, there exists a translation model trained on different views of the bilingual corpus (different translation direction or decoding order as in ST, BT and TA) or different parameterization (the SMT-like model in DA4Low), which can guarantee the equivariance and smoothness properties of $(\tilde{\mathrm{x}}, \tilde{\mathrm{y}})$. That is, the original model is trained on the outputs from another translation model, which matches the knowledge distillation paradigm (Hinton et al., 2015; Kim and Rush, 2016; Furlanello et al., 2018). So we call them knowledge distillation (kn) based DA methods.

Besides the above DA methods, most other DA methods can be a variant of one of them. For kn bsaed methods, reconstruction (Cheng et al., 2016) is like BT when used on monilingual data with instance reweighting; and when the augmentation process is seen as a stage in iterative co-training between the target model and the augmentation model, dual learning (Xia et al., 2016), and joint training (Zhang et al., 2018) can be unified.

B Translation Tasks and Training Details

Table 5 shows the statistics of the three standard benchmarks we rely on, with the IWSLT corpus for training both translation directions so as to obtain four translation tasks. We choose corpora with different sizes: 0.22M, 1M and 4.5M. All the corpora are publicly available from their websites. [4] [5] [6] To note that we choose the datum2017 corpus as a subset of the Zh⇒En corpus for constructing the medium sized translation task. All the data is pre-processed with Byte Pair Encoding (Sennrich et al., 2016b) by jointly learning the source and target vocabulary. [7]

Table 6 shows the hyper-parameters of training on each translation tasks. In Section 2.1, we have identified two factors to control the effect of learning from $\mathcal{T}$ and $\mathcal{A}$. We conduct experiment among DA methods where the two factors are the same or at least similar. According to the categorization discussed in Appx. A. The dn based methods are doing online augmentation so the interpolation coefficient α equals to the probability an instance is to be augmented. This quantity is derived to be around 0.6. So for ST, TA, BT, we use beam search to augment every instance in the train with the top one decoded instance, so that the α is around 0.5 which is comparable to dn based methods. We do not consider DA4Low in our experiment since the α is around 0.05 which are far from 0.6.

C Kendall's Coefficient of Concordance

Kendalls coefficient of concordance (Mazurek, 2011) is computed through the following formula:

$$W = \frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_i^n X_i)^2}{n}}{\frac{1}{12} \cdot k^2 \cdot (n^3 - n)}, \tag{9}$$

where k is the number of rankings and n the number of objects. In our setting, k is 4 corresponding to the four translation tasks and n is 6 corresponding to the 5 DA methods plus the baseline.

D Measure Binned Avg. Freq. Statistics

This appendix section demonstrates the measure (input sensitivity or prediction margin) binned statistics of our two measures on the other translation tasks in Figure 3 and 4 respectively. The x-axis is one of the measures and the y-axis is the average token frequency within that bin. For the both the measure, we can also see similar trends that both of the measures are improved largely for low frequency tokens instead of high-frequency tokens. Actually, we find that most of the high frequency tokens sacrifice their sensitivity or margin as a trade-off with low frequency tokens.

[4] https://wit3.fbk.eu/mt.php?release=2017-01-trnted

[5] http://nlp.nju.edu.cn/cwmt-wmt/

[6] http://www.statmt.org/wmt19/translation-task.html

[7] https://github.com/rsennrich/subword-nmt

Method		Generative story of $\tilde{x}, \tilde{y}$ with q_{AUG}		src/tgt	dn/kn	o/f
RAML	Norouzi et al. (2016)	$\tilde{x} = x$;	$\tilde{y} \sim p_{\tilde{y}\|y}$	tgt	dn	o
SO	Wang et al. (2018)	$\tilde{x} \sim p_{\tilde{x}\|x}$;	$\tilde{y} \sim p_{\tilde{y}\|y}$	both	dn	o
ST	Zhang and Zong (2016)	$\tilde{x} = x$;	$\tilde{y} \sim p_{\theta',l2r}(\cdot\|\tilde{x})$	tgt	kn	f
TA	Zhang et al. (2019)	$\tilde{x} = x$;	$\tilde{y} \sim p_{\theta',r2l}(\cdot\|\tilde{x})$	tgt	kn	f
BT	Sennrich et al. (2016a)	$\tilde{y} = y$;	$\tilde{x} \sim p_{\theta',l2r}(\cdot\|\tilde{y})$	src	kn	f
DA4Low	Fadaee et al. (2017)	$\tilde{y} \sim p_{\gamma,lm}(\cdot\|y)$;	$\tilde{x} \sim p_{\theta',smt}(\cdot\|\tilde{y})$	both	kn	f

Table 4: Categorization of various DA methods according to the augmentation distribution. Note that each DA method is given a name abbreviation (RAML: Reward Augmented MLE, SO: Switchout, ST: Self-Training, TA: Target-side Agreement regularization, BT: Back-Translation). The **generative story** column describes specific choices of q_{AUG} and the generation of an augmented instance $(\tilde{x}, \tilde{y})$. Here the sampling process $x, y \sim \hat{p}$ is omitted since it exists in every DA method. $p_{\theta'}$ is another NMT model trained with different translation direction (conditioned on src/tgt) or different decoding order (l2r or r2l). p_{γ} is might be a well trained bidirectional language model from which we can sample and replace sub-spans in y. The **src/tgt** column shows the side of language a DA method augments. The **dn/kn** column classifies a DA method into data noising based or knowledge distillation based. The **o/f** column classifies a DA method into online or offline augmentation.

Tasks	train	dev	test	BPE merge no.	src vocab	tgt vocab
Fr⇔En	218878	9948	9487	10000	11981	9840
Zh⇒En	998668	3002	3981	60000	46953	37071
En⇒De	4542403	3000	3003	40000	39996	39996

Table 5: Corpus statistics of the four translation tasks.

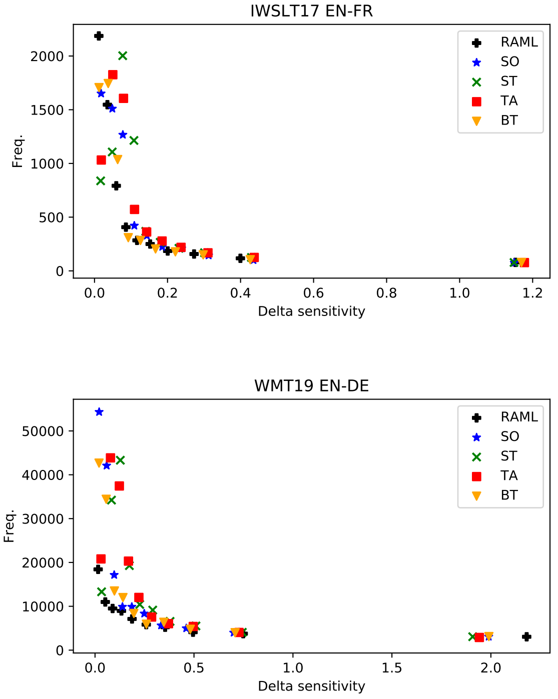

Figure 3: Δ sensitivity binned average token frequency statistics on IWSLT17 Fr⇒En (0.22M) and WMT19 En⇒De (4.5M).

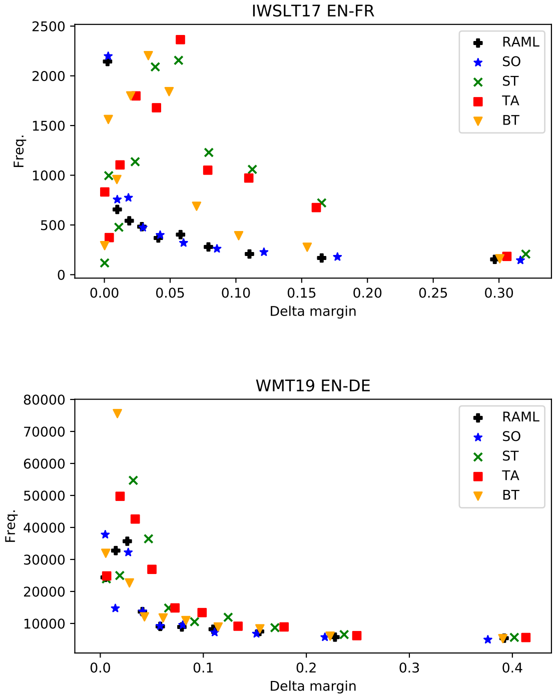

Figure 4: Δ margin binned average token frequency statistics on IWSLT17 Fr⇒En (0.22M) and WMT19 En⇒De (4.5M).

Tasks	n_{layers}	n_{head}	d_{model}	d_{inner}	sched.	n_{warmup}	n_{epoch}	init. lr
Fr⇒En	2	4	256	512	Switchout	-	80	0.001
En⇒Fr	2	4	256	512	Switchout	-	80	0.001
Zh⇒En	6	8	512	2048	inverse_sqrt	4000	30	0.0007
En⇒De	6	8	512	2048	inverse_sqrt	4000	35	0.0007

Table 6: Hyper-parameters for the model and the training.